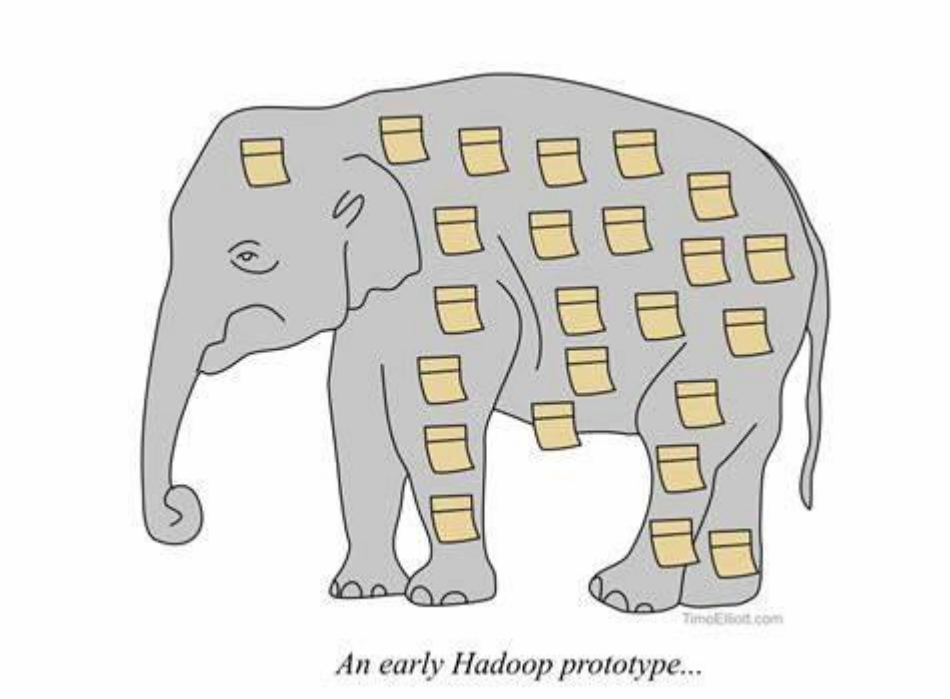


An early Hadoop prototype...

ASSIGNMENT 1 HADOOP

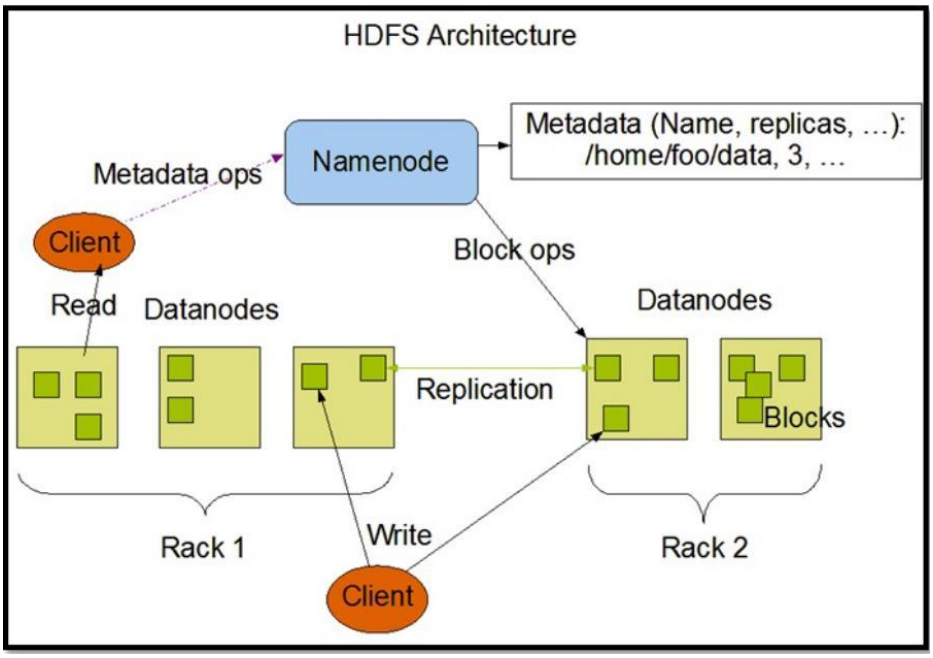
ITRI 623

ENRICO
DREYER
31210783

Table of Contents

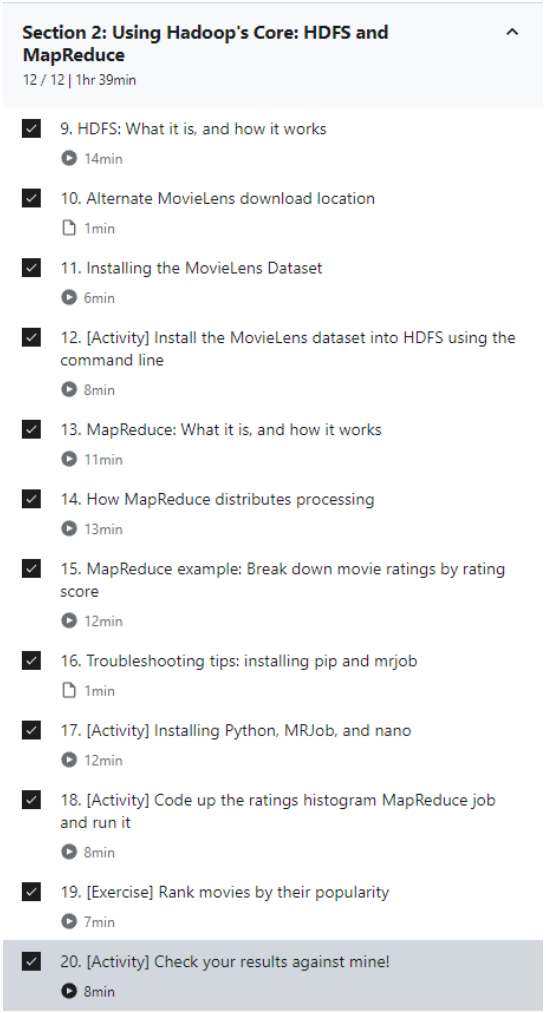
Introduction

For this assignment we were asked to make a video of no longer than 2 minutes on any two sections of the Hadoop Udemy Course. For the first Assignment I decided to go with the basics of Hadoop, this being HDFS and MapReduce.



The following section of the document will discuss the notes of my video

Proof



My Random notes of the course:

The Ultimate Hands-On Hadoop: Tame your Big Data!

HDFS Architecture - HDFS works by having a main 'NameNode' and multiple 'data nodes' on the same commodity hardware cluster. All nodes are almost always grouped in the same physical rack in the data center. Then data is broken down into different 'blocks' that are shered among the different data nodes for storage. Blocks are also make an exact copy of across nodes to reduce the chances of failure. (Nicolas, 2014).

Putty cmd - PuTTY (/ˈpʌti/) is a free and open-source terminal emulator, serial console and network file transfer application.

hadoop fs -ls For a directory: returns the list of directories and file, for a file: returns the statistics on the file.

hadoop fs -lsr: recursively listing the files and directories under the specific folders.

Example: Hadoop fs -ls / or hadoop fs -lsr

MapReduce is for writing applications; it is a programming model that can process Big Data in a parallel manner on more than one node at a time. MapReduce has the ability to provide analytical capabilities for analyzing big amounts of complex data. (Phoenixnap, 2020)

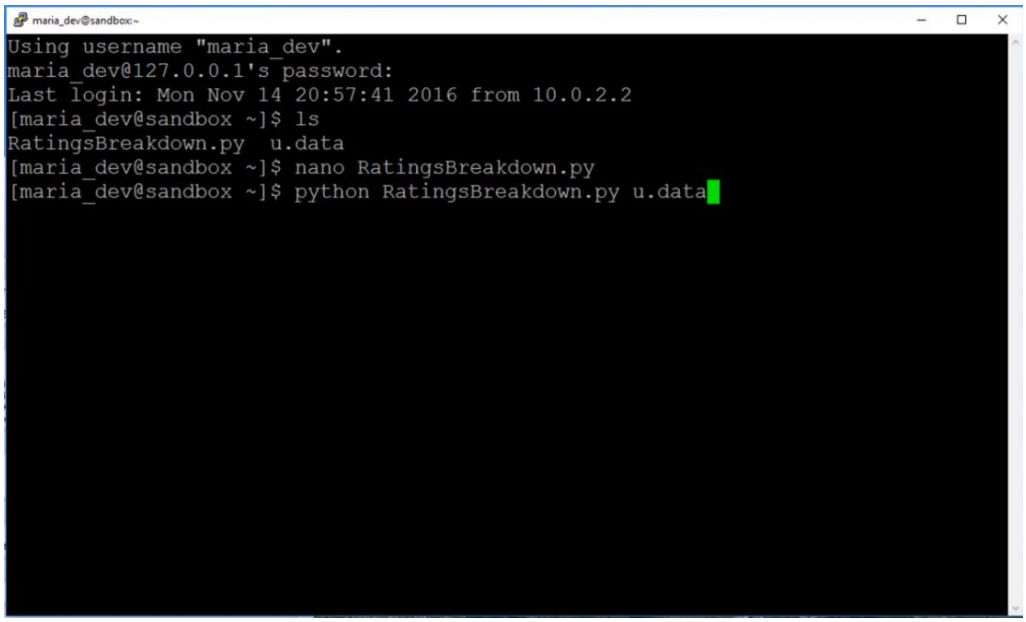
Key-Value Pair (KVP) Mapping is the process of linking the key to its associated value. While mapping, if the key 'firstName' is associated with a value 'Bugs', it means that the array maps the 'Bugs' to key 'firstName' (Servicenow, 2021).

key	value
firstName	Bugs
lastName	Bunny
location	Earth

Shuffle and Sort – The shuffle phase in Hadoop changes the map output from Mapper to a Reducer in MapReduce. The sort phase in MapReduce covers the sorting and merging of map outputs. The data from the mapper are grouped by the key, then split among reducers, and then sorted by the key. Every reducer gets all the values associated with the same key (DataFair, 2021).

Hadoop Streaming – The reducer and the mapper are executables that can read the input line by line and emit the output to line by line. The utility can then create a MapReduce job, submit the job to an appropriate cluster, and monitor the progress of the job until it completes.

Practical:



```
from mrjob.job import MRJob
from mrjob.step import MRStep

class RatingsBreakdown(MRJob):
    def steps(self):
        return [
            MRStep(mapper=self.mapper_get_ratings,
                   reducer=self.reducer_count_ratings),
            MRStep(reducer=self.reducer_sorted_output)
        ]

    def mapper_get_ratings(self, _, line):
        (userID, movieID, rating, timestamp) = line.split('\t')
        yield movieID, 1

    def reducer_count_ratings(self, key, values):
        yield str(sum(values)).zfill(5), key

    def reducer_sorted_output(self, count, movies):
        for movie in movies:
            yield movie, count

if __name__ == '__main__':
    RatingsBreakdown.run()
```

Map – reduce – sort

```
from mrjob.job import MRJob
from mrjob.step import MRStep

class RatingsBreakdown(MRJob):
    def steps(self):
        return [
            MRStep(mapper=self.mapper_get_ratings,
                   reducer=self.reducer_count_ratings)
        ]

    def mapper_get_ratings(self, _, line):
        (userID, movieID, rating, timestamp) = line.split('\t')
        yield rating, 1

    def reducer_count_ratings(self, key, values):
        yield key, sum(values)

if __name__ == '__main__':
    RatingsBreakdown.run()
```

This is specific to counting the movies for each rating.

References

DataFair. (2021). Shuffling and Sorting in Hadoop MapReduce. https://data-flair.training/blogs/shuffling-and-sorting-in-hadoop/

Nicolas. (2014). WHAT IS HADOOP AND HOW DOES IT WORK? https://dataconomy.com/2014/02/hadoop-what-how-introduction/#:~:text=The%20way%20HDFS%20works%20is%20by%20having%20a,distributed%20among%20the%20various%20data%20nodes%20for%20storage.

Phoenixnap. (2020). What is Hadoop Mapreduce and How Does it Work. https://phoenixnap.com/kb/hadoop-

mapreduce#:~:text=MapReduce%20is%20a%20processing%20module%20in%20the%20Apache,use%20low-cost%20consumer%20hardware%20to%20handle%20your%20data.
Servicenow. (2021). Create a key-value pair mapping. https://docs.servicenow.com/bundle/rome-now-intelligence/page/use/reporting/task/t_CreateKeyValuePairMapping.html